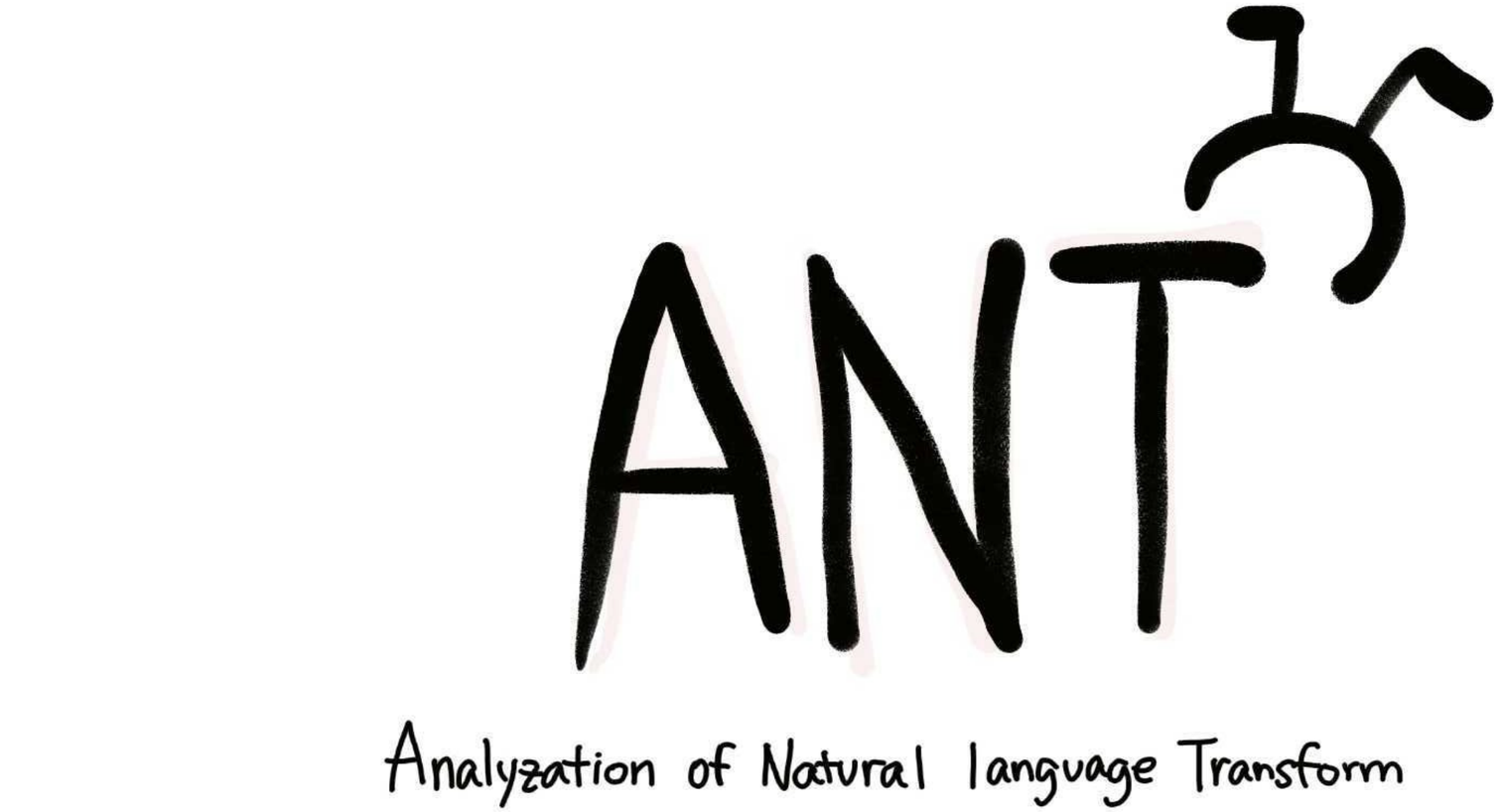
Analyzation of Natural language Transform

Team 1조 한지애, 김혜인, 박장선, 이은지, 최현호, 홍석영

목차

01. 프로젝트 개요

01 프로젝트명



이해 vs 생성

01 프로젝트 주제 및 선정 배경

이해 vs 생성

01 프로젝트 주제 및 선정 배경

> 국내 뉴스 기사를 요약하고 요약된 기사를 한/영 번역 모델을 이용하여 영문으로 변환함으로써 (다문화) 사용자가 최신/인기 기사를 쉽게 접근할 수 있는 뉴스 서비스 개발

> 요약, 한국어- 영어 번역 모델 개발

> Pretrained Model 사용 / Fine- Tuning 이해
- 생성 모델 한계 : 다량의 데이터, 문장 단위로 tagging 되어있는 pair 데이터 필요
자체적으로 학습에 필요한 데이터 구축이 어려움

01 프로젝트 개요

> 요약, 번역, 요약과 번역을 동시에 진행하는 웹 서비스 개발
- 요약과 변역 모두 글의 이해를 돕는 기능
- 기사 전체 본문을 간략한 문장으로 변환 후, 영어로 번역해서 전달
- 영어를 사용하는 외국인이 한국어 글을 읽을 때 이해를 돕는 효과

<모델>

- 요약
- 번역
- 요약 및 번역

-> Transformer 모델
Python, Pytorch

<웹>

- Django (HTML)

01 프로젝트 팀 구성 및 역할

팀장 한지애	> 데이터 전처리 > 요약 모델 개발 > Web 디자인
이은지	> 데이터 전처리 > 번역 모델 개발 > PPT 제작
최현호	> 데이터 전처리 > 번역 모델 개발
박장선	> 데이터 전처리 > 요약모델 개발 > git 구축 , Web
홍석영	> 데이터 전처리 > 요약 모델 개발
김혜인	> 데이터 전처리 > 번역 모델 개발 > Web

01 활용장비 및 재료(개발환경)

개발 환경

- language: python 3.10.11
- tool: visual_studio_code,colab,git_hub
- NVIDIA A100-SXM4-40GB

library

- accelerate : 0.20.1
- pytorch : 2.0.1+cu118
- transformers : 4.33
- pytorch-lightning : 2.0.9
- safetensors : 0.3.1
- sacremoses : 0.3.1

커스텀 모듈

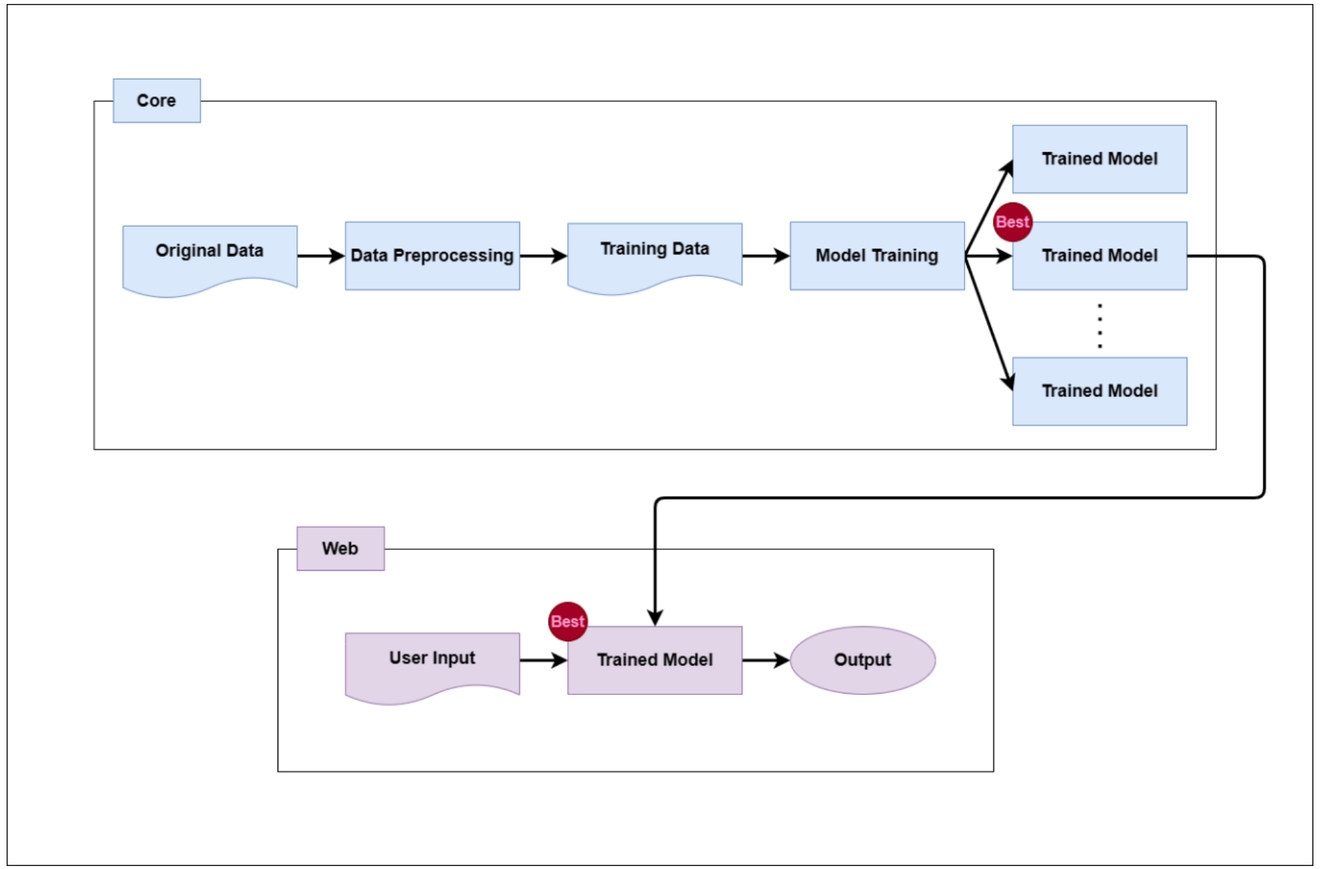
- dataset_maker v1.2,
- load_modeler v1.0

컴퓨팅 단위

Google Colab

- GPU RAM 40GB

01 프로젝트 구조


Core
Original Data
Data Preprocessing
Training Data
Model Training
Trained Model
Best
Trained Model
Trained Model
Web
User Input
Best
Trained Model
Output

02. 프로젝트 수행 절차 및 방법

02 프로젝트 수행 절차_과정

구분	기간	활동	비고
사전 기획	8/2(수) ~ 8/13(금)	> 프로젝트 기획 및 주제 선정 > 기획안 작성	> 아이디어 선정
데이터 수집	8/13(월) ~ 8/16(금)	> 필요 데이터 및 수집 절차 정의 > 외부 데이터 수집	> 협약 기업 데이터 협조
데이터 전처리	8/14(월) ~ 8/16(수)	> 데이터 정제 및 정규화	-
모델링	8/16(월) ~ 9/20(수)	> 모형 구현	> 팀별 중간보고 실시
서비스구축	8/13(월) ~ 9/20(수)	> 장고 설계 > 장고 플랫폼 구현	> 최적화 , 오류 수정
총 개발 기간	8/20(월) ~ 9/20(수) (총 7주)	-	-

03-1 데이터셋 소개

03-1 데이터셋_요약

데이터 종류	기사 원문	요약문
뉴스 기사	21600	21600

> 원천 데이터 : AI Hub 요약문 및 레포트 생성 데이터

> 학습 데이터 : 뉴스기사 json 파일 내 “passage”, “summary1”을 추출하여 .tsv 파일로 변환 후 랜덤으로 10000개 추출

- 데이터 전처리 : 불필요한 단어(재판매 및 DB 금지), 소괄호(© News1|※) , 광고, “\n” 등 ‘passage’데이터 내 학습에 방해가 되는 문자 제거

(aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=582)

03-1 데이터셋_요약 전처리

기사 원문 전처리 전

회사원 주동형(30) 씨는 식당에서 밥값을 낼 때 신용카드 대신 휴대전화를 꺼낸다. <-엔터
카드 단말기에 휴대전화를 접촉하는 '삼성페이'를 이용한다. 집에서는 '네이버페이'로 카드번호를 입력하지 않고도 온라인 쇼핑을 즐긴다. 이 정도는 이미 일상이 됐다. 새해에는 아예 지갑 없이도 종일 살 수 있을까. 좀 더 쉽고, 빨라질 건 확실해 보인다. 2018년 국내 모바일 간편결제 가입자(중복 포함)는 약 1억7000만명. 이들이 한 해 간편결제로 이용한 금액은 약 80조원에 달한다. 한국금융투자자보호재단이 지난해 만 25~64세 성인 남녀 2530명을 상대로 한 조사 결과, 56. <- 소수점
8%가 간편결제 서비스를 이용한다고 답했다. 이커머스(e-commerce)시장장이 중심에 있다.

<- 띄어쓰기 2개

전처리 후

회사원 주동형(30) 씨는 식당에서 밥값을 낼 때 신용카드 대신 휴대전화를 꺼낸다.카드 단말기에 휴대전화를 접촉하는 '삼성페이'를 이용한다.집에서는 '네이버페이'로 카드번호를 입력하지 않고도 온라인 쇼핑을 즐긴다.이 정도는 이미 일상이 됐다.새해에는 아예 지갑 없이도 종일 살 수 있을까.좀 더 쉽고, 빨라질 건 확실해 보인다.2018년 국내 모바일 간편결제 가입자(중복 포함)는 약 1억7000만명.이들이 한 해 간편결제로 이용한 금액은 약 80조원에 달한다.한국금융투자자보호재단이 지난해 만 25~64세 성인 남녀 2530명을 상대로 한 조사 결과, 56.8%가 간편결제 서비스를 이용한다고 답했다.이커머스(e-commerce)시장장이 중심에 있다.

03-1 데이터셋_번역

데이터 종류	원문(한국어)	번역문(영어)
뉴스 기사	80만	80만

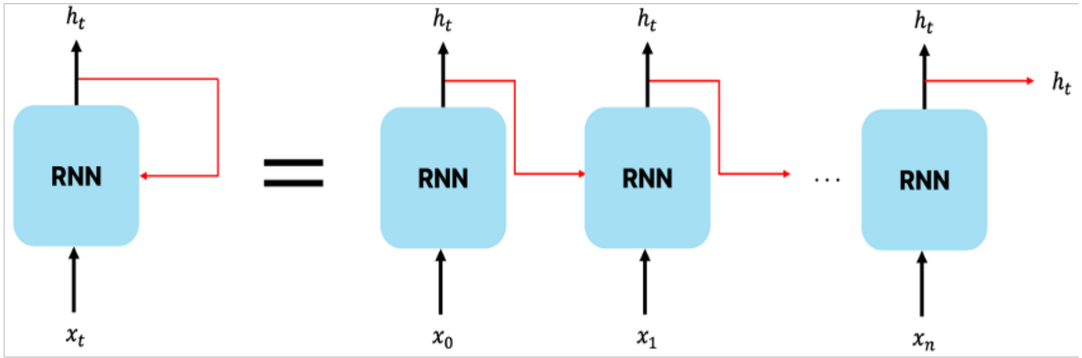
> 원천 데이터 : AI Hub 한국어-영어 번역(병렬) 말뭉치
> 학습 데이터 : 뉴스 분야의 80만 문장 .csv 변환 후 랜덤으로 100000개 사용

((https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=126)

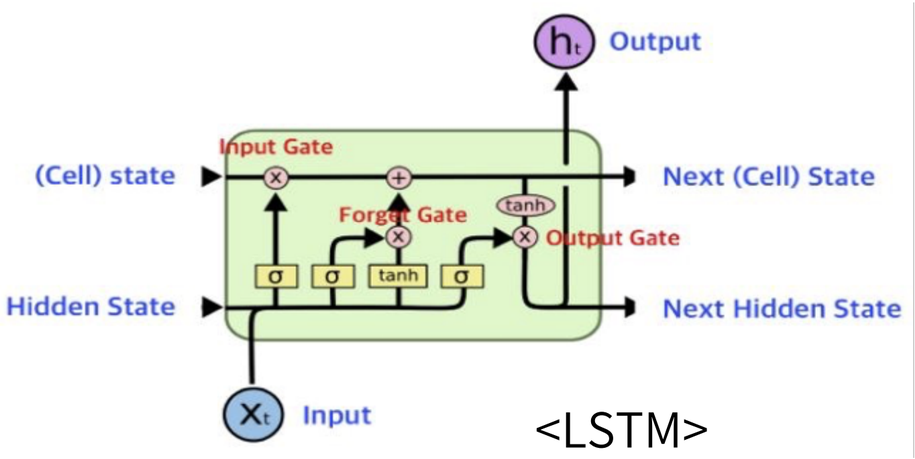
*순수 문자열이 대부분인 번역 데이터는 전처리 진행 X

03-2 모델 소개

03-2 RNN & LSTM



< RNN>

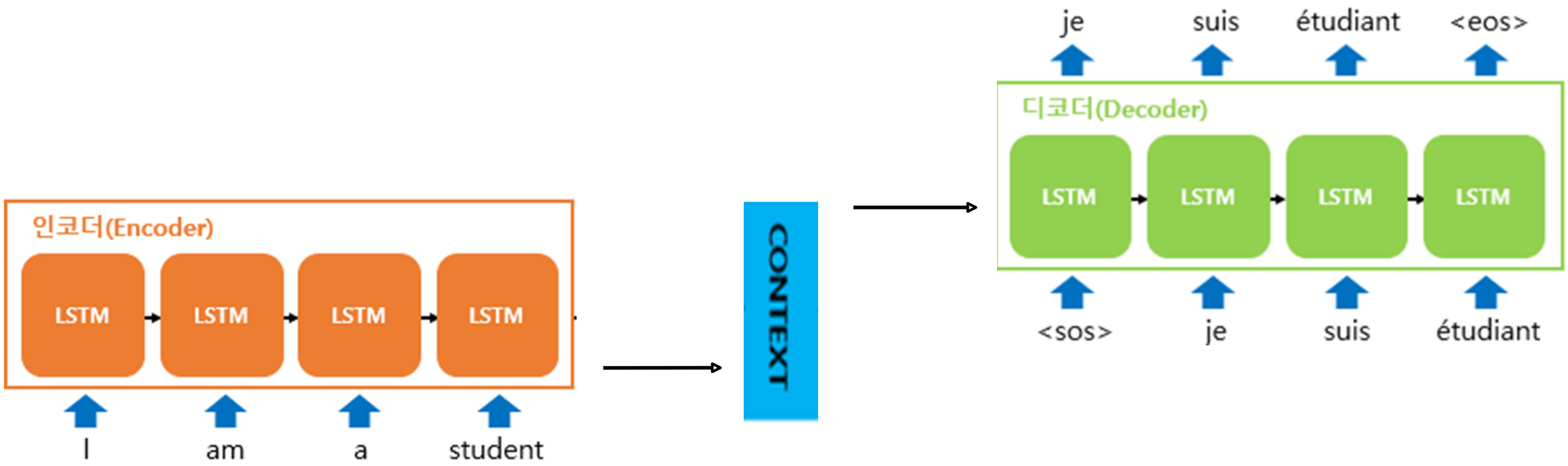


<LSTM>

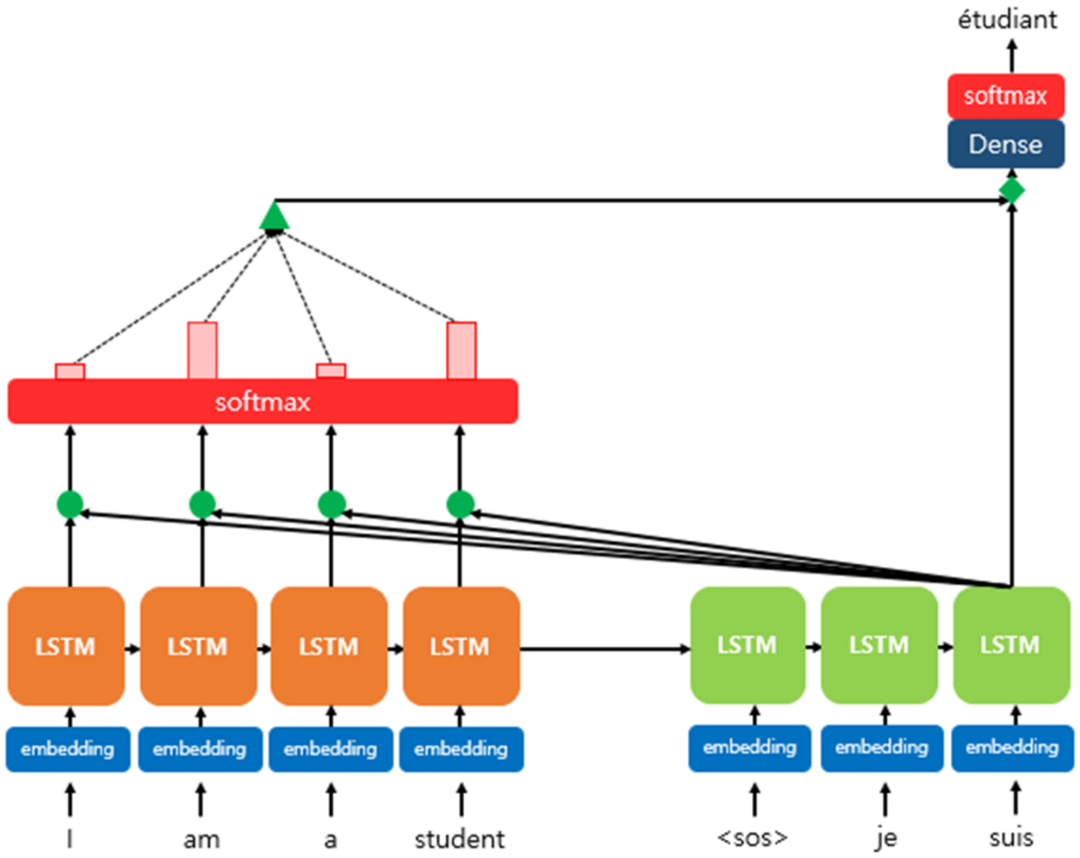
> 단점

- 장기 의존성 문제(the problem of Long-Term Dependencies)
- Vanishing Gradient
- 병렬 처리가 불가능하여 순차적인 학습으로 진행 속도가 느림

03-2 SeqtoSeq



03-2 Attention



> 해당 쿼리에서
> 예측해야 할 단어(토큰)과
> 연관이 있는 입력 단어(토큰) 부분을
> 좀 더 집중 *Attention!*

03-2 Transformer

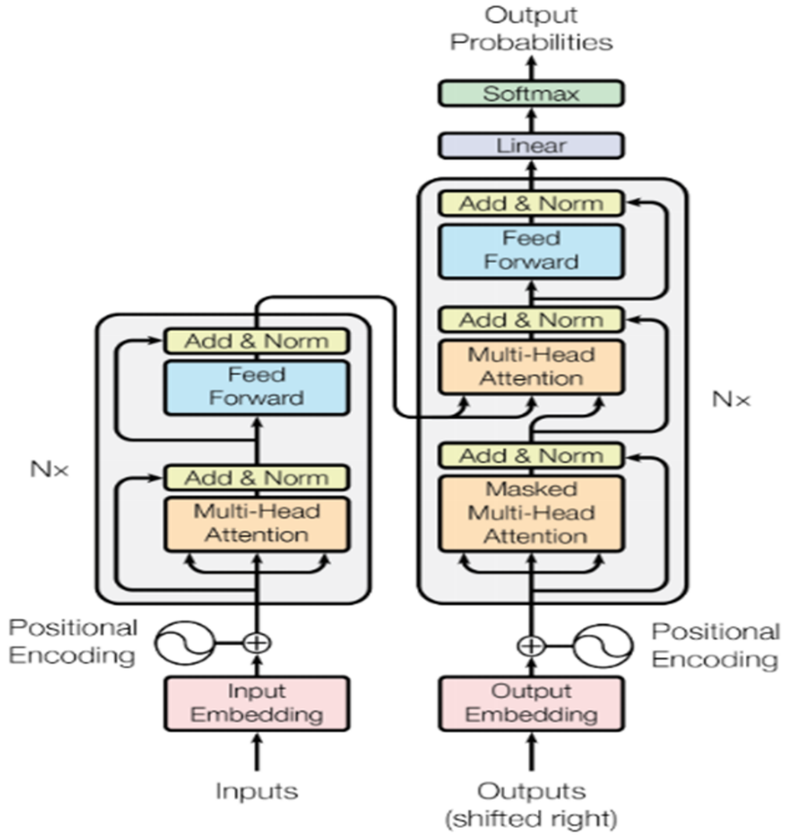
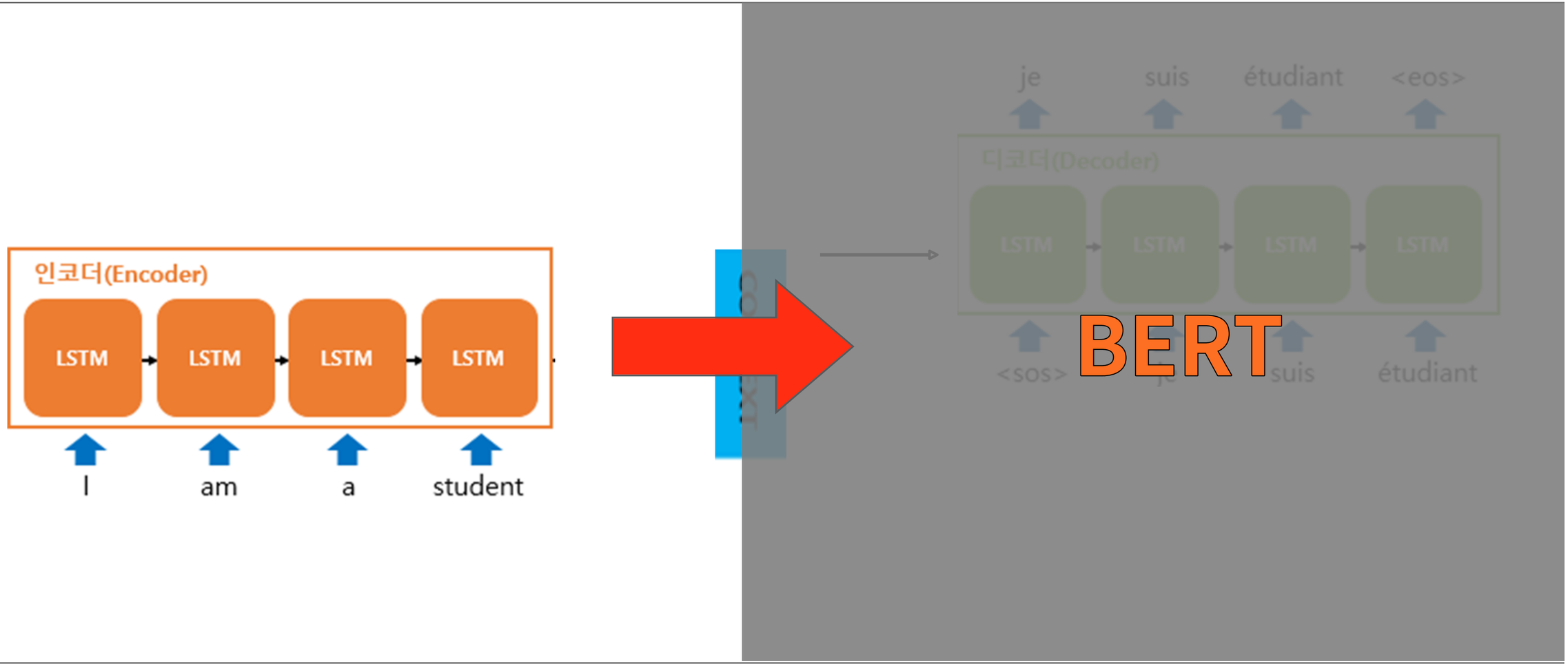


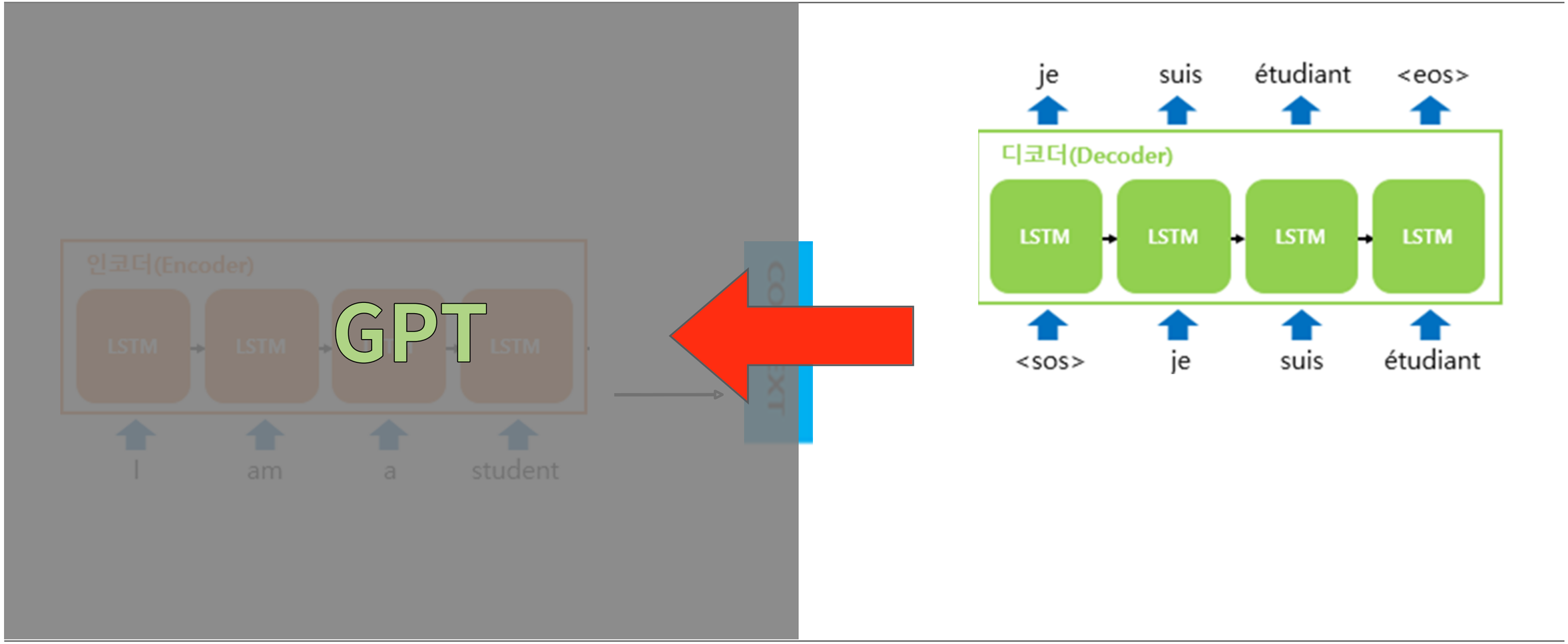
Figure 1: The Transformer - model architecture.

Attention is All you need !

03-2 BART


인코더(Encoder)
LSTM
LSTM
LSTM
LSTM
I
am
a
student
je
suis
étudiant
<eos>
디코더(Decoder)
LSTM
LSTM
LSTM
LSTM
<sos>
suis
étudiant
BERT

03-2 BART


je
suis
étudiant
<eos>
디코더(Decoder)
LSTM
LSTM
LSTM
LSTM
<sos>
je
suis
étudiant
GPT
인코더(Encoder)
LSTM
LSTM
LSTM
I
am
a
student

03-2 BART(Bidirectional Auto-Regressive Transformer)

사전학습된 생성 모델
-> Transformer를 기반으로 하는 **BERT + GPT = BART**

> 요약 : KoBART 사용
- 한국어에 특화된 요약 모델 KoBART (digit82/kobart-summarization)

> 번역 : MarianModel 사용 ("Helsinki-NLP/opus-mt-ko-en")
- 번역에 특화된 BART 기반 모델

공통점 : 많은 양의 데이터 학습 필요, BartForConditionalGeneration

04. 프로젝트 수행 결과

04. 성능 지표

Summaray Model	Data Size	Batch Size	Epoch	Time	Accuracy	Translation Model	Data Size	Batch Size	Epoch	Time	Accuracy
digit82/kobart-summarization	100	10	100	10M	99%	Hellsinki - NLP/opus-mt-ko-en	100	10	100	2M	69%
	200	10	200	12M	95%		200	10	200	10M	99%
	1000	10	100	1H35M	99.7%		1000	10	1000	1H	92%
	10000	20	33	6H32M	98.5%		10000	10	150	12H30M	94%

04. 경량화 시도

```
  (decoder): BartDecoder(
    (embed_tokens): Embedding(30000, 768, padding_idx=3)
    (embed_positions): BartLearnedPositionalEmbedding(1028, 768)
    (layers): ModuleList(
      (0-5): 6 x BartDecoderLayer(
        (self_attn): BartAttention(
          (k_proj): Linear(in_features=768, out_features=768, bias=True)
          (v_proj): Linear(in_features=768, out_features=768, bias=True)
          (q_proj): Linear(in_features=768, out_features=768, bias=True)
          (out_proj): Linear(in_features=768, out_features=768, bias=True)
        )
        (activation_fn): GELUActivation()
        (self_attn_layer_norm): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
        (encoder_attn): BartAttention(
          (k_proj): Linear(in_features=768, out_features=768, bias=True)
          (v_proj): Linear(in_features=768, out_features=768, bias=True)
          (q_proj): Linear(in_features=768, out_features=768, bias=True)
          (out_proj): Linear(in_features=768, out_features=768, bias=True)
        )
        (encoder_attn_layer_norm): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
        (fc1): Linear(in_features=768, out_features=3072, bias=True)
        (fc2): Linear(in_features=3072, out_features=768, bias=True)
        (final_layer_norm): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
      )
    )
    (layernorm_embedding): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  )
)
(lm_head): Linear(in_features=768, out_features=30000, bias=False)
```

04. 요약_학습전후 비교

원문

- 신종코로나바이러스 감염증(코로나19)으로 소득이 급감했는데 신용대출 만기가 다가온다면 어떻게 해야 할까. 연체에 빠지는 것이 두려워서 급한 불을 끄기 위해 자칫 고금리로 급전을 끌어써야 겠다는 위험한 생각을 하기가 쉽다. 이런 사람들을 위해 금융권이 대책을 내놨다.코로나19 피해를 입은 개인대출자들을 위한 가계대출 원금 상환유예 조치가 29일부터 전 금융권에서 시행된다. 대상은?코로나19 관련 대출원금 상환유예 프로그램의 정식 명칭은 '프리워크아웃'이다.신용회복위원회의 워크아웃 대상(연체기간 90일 이상)이 아닌, 연체 직전 또는 단기 연체 채무자를 위한 채무조정 프로그램이다. 기존에도 프리워크아웃 제도는 있었지만 코로나19로 소득 감소 등 피해를 입은 채무자를 위해 문턱을 낮춰줬다. 대상은 코로나19피해로 소득이 감소해서 생계비(복지부 고시 기준중위소득의 75%)를 빼고 나면 빚을 갚기가 어려운 채무자다. 가계대출 중에서도 신용대출과 보증부 서민금융대출(햇살론·사잇돌대출 등)만 대상이다. 주택담보대출처럼 담보가 있는 대출은 제외된다.카드사용대금, 현금서비스, 오토론, 보험약관대출도 제외다.기본적으로 방안이 발표되기 전인 4월 8일 이전에 체결된 대출계약에 이를 적용한다.4월 9일 이후에 증액 또는 신규대출이 이뤄진 경우엔 금융회사가 판단해 대상에서 제외할 수 있다.신청은 어떻게?상환유예가 필요한 대출의 해당 금융회사가 1곳이라면 그 금융사에 신청하면 된다.2곳 이상의 금융회사에 신청해야 한다면 신용회복위원회를 통해 한번에 신청할 수 있다.대상이 된다고 해서 아무 때나 신청할 수 있는 건 아니다.개별 금융회사에 신청하는 경우엔 원금을 갚아야 하는 예정일이 1개월 미만으로 남았을 때만 신청할 수 있다.이와 달리 신용회복위원회엔 원금 상환예정일과 관계없이 신청할 수 있다.은행, 보험, 저축은행, 카드, 캐피탈, 상호금융(농협·수협·산림조합·새마을금고)의 약 3700개 금융사가 참여했다.신청기간은 이달 29일부터 올해 12월 31일까지다.다만 두 인터넷은행(카카오뱅크, 케이뱅크)은 다음달 7일부터 신청할 수 있다. 신청을 위한 수수료 부담은 없다. 혜택은 무엇?프리워크아웃 대상자엔 대출원금 상환을 6~12개월 미뤄준다. 예컨대 올해 5월에 만기가 돌아오는 일시상환대출이라면 만기를 올해 11월~내년 5월 사이로 연장할 수 있다.

정답 요약문

코로나19 피해로 소득이 감소해 생계비를 빼고 나면 빚을 갚기 어려운 채무자를 위해 가계대출 원금 상환유예 조치가 전 금융권에서 시행되며 신용대출과 서민금융대출만 대상이다.

학습 전 요약문

신종코로나바이러스 감염증으로 소득이 급감했는데 신용대출 만기가 다가온다면 어떻게 해야 할까.

학습 후 요약문

코로나19 피해로 소득이 감소해 생계비를 빼고 나면 빚을 갚기 어려운 채무자를 위해 가계대출 원금 상환유예 조치가 전 금융권에서 시행되며 신용대출과 서민금융대출만 대상이다.

04. 번역_학습전후 비교

원본 요약문

1:1 매칭과 전략 매칭 형식으로 진행된 이번 비즈니스 상담회에는 한국 31개사 및 중국 유쿠(YOUKU), 아이치이(IQIYI) 등 150여 개사가 참여해 이 중 웹툰 기업은 상담횟수242건을 달성했다.

정답 번역문

The business counseling session, which was conducted in the form of 1:1 matching and strategy matching, involved 31 South Korean companies, including YOUKU of China and IQIYI of China, and among them, Webtoon companies achieved 242 cases of counseling.

학습 전 번역문

The building was meeting in the name of 1: 1 and searching, involuntarily observing 31 Southeran associations, including YOKU of China and IQI, where they came from, 242 years ago.

학습 후 번역문

The best Counseling integration was making one-on-one meet, introduced 31 South Korean companies, and inflowing YOUKU of China and IQIYI, of which webtoon companies achieved 242 consultations.

04. 요약 및 번역_학습 후

원문

- 신종코로나바이러스 감염증(코로나19)으로 소득이 급감했는데 신용대출 만기가 다가온다면 어떻게 해야 할까. 연체에 빠지는 것이 두려워서 급한 불을 끄기 위해 자칫 고금리로 급전을 끌어써야 겠다는 위험한 생각을 하기가 쉽다. 이런 사람들을 위해 금융권이 대책을 내놨다.코로나19 피해를 입은 개인대출자들을 위한 가계대출 원금 상환유예 조치가 29일부터 전 금융권에서 시행된다. 대상은?코로나19 관련 대출원금 상환유예 프로그램의 정식 명칭은 '프리워크아웃'이다.신용회복위원회의 워크아웃 대상(연체기간 90일 이상)이 아닌, 연체 직전 또는 단기 연체 채무자를 위한 채무조정 프로그램이다. 기존에도 프리워크아웃 제도는 있었지만 코로나19로 소득 감소 등 피해를 입은 채무자를 위해 문턱을 낮춰줬다. 대상은 코로나19피해로 소득이 감소해서 생계비(복지부 고시 기준중위소득의 75%)를 빼고 나면 빚을 갚기가 어려운 채무자다. 가계대출 중에서도 신용대출과 보증부 서민금융대출(햇살론·사잇돌대출 등)만 대상이다. 주택담보대출처럼 담보가 있는 대출은 제외된다.카드사용대금, 현금서비스, 오토론, 보험약관대출도 제외다.기본적으로 방안이 발표되기 전인 4월 8일 이전에 체결된 대출계약에 이를 적용한다.4월 9일 이후에 증액 또는 신규대출이 이뤄진 경우엔 금융회사가 판단해 대상에서 제외할 수 있다.신청은 어떻게?상환유예가 필요한 대출의 해당 금융회사가 1곳이라면 그 금융사에 신청하면 된다.2곳 이상의 금융회사에 신청해야 한다면 신용회복위원회를 통해 한번에 신청할 수 있다.대상이 된다고 해서 아무 때나 신청할 수 있는 건 아니다.개별 금융회사에 신청하는 경우엔 원금을 갚아야 하는 예정일이 1개월 미만으로 남았을 때만 신청할 수 있다.이와 달리 신용회복위원회엔 원금 상환예정일과 관계없이 신청할 수 있다.은행, 보험, 저축은행, 카드, 캐피탈, 상호금융(농협·수협·산림조합·새마을금고)의 약 3700개 금융사가 참여했다.신청기간은 이달 29일부터 올해 12월 31일까지다.다만 두 인터넷은행(카카오뱅크, 케이뱅크)은 다음달 7일부터 신청할 수 있다. 신청을 위한 수수료 부담은 없다. 혜택은 무엇?프리워크아웃 대상자엔 대출원금 상환을 6~12개월 미뤄준다. 예컨대 올해 5월에 만기가 돌아오는 일시상환대출이라면 만기를 올해 11월~내년 5월 사이로 연장할 수 있다.

요약문

코로나19 피해로 소득이 감소해 생계비를 빼고 나면 빚을 갚기 어려운 채무자를 위해 가계대출 원금 상환 유예 조치가 전 금융권에서 시행되며 신용대출과 서민금융대출만 대상이다.

번역문

The suspension of repayment of the principal of household loans will be implemented in all financial sectors for debtors who are unable to pay off their debts after their income decreases due to COVID-19 damage, and only credit loans and financial loans for the working class are covered.

파파고

Deferral measures to repay the principal of household loans will be implemented in all financial sectors for debtors who are unable to pay off their debts after their income decreases due to the damage of COVID-19, and only credit loans and low-income financial loans will be covered.

04. 요약_opentest

기사 원문

미국에서 스타벅스가 과일 이름이 들어간 음료에 과일이 들어가지 않았다는 이유로 집단소송을 치르게 됐습니다.

현지시간 18일 로이터 통신 등에 따르면 미국 뉴욕 남부지방법원은 이날 "소비자 대부분이 스타벅스 과일 음료에 실제로 과일이 포함됐다고 생각할 것"이라면서 스타벅스의 소송 기각 요청을 받아들이지 않았습니다.

지난해 8월 뉴욕과 캘리포니아 출신 원고 2명은 스타벅스 과일 음료 '망고 드래곤푸르트', '파인애플 패션푸르트', '스트로베리 아사이 레모네이드 리프레셔' 등에 실제로 망고나 패션푸르트, 아사이가 없어 스타벅스가 소비자 보호법을 위반했다며 소송을 냈습니다.

원고가 주장한 피해 집단에 대한 배상 금액은 최소 500만 달러(약 66억원)로 전해졌습니다.

이에 스타벅스는 소송이 기각돼야 한다면서 해당 제품명은 음료 성분이 아닌 맛을 설명한 것이라고 주장했습니다.

그러나 존 크로넌 담당 판사는 '아이스 말차 라떼'에는 말차가 실제로 들어가는 것처럼 일부 스타벅스 음료 이름이 성분 이름을 따서 지어졌다는 점을 고려해 소비자가 해당 과일 음료에도 과일이 포함됐다고 생각할 수 있다고 판단했습니다.

크로넌 판사는 스타벅스가 소비자를 속이려 하거나 부당이득을 취하려 한 것은 아니라고 봤습니다.

스타벅스 성명을 통해 "소송의 주장이 부정확하고 타당성이 없다"면서 "우리는 이에 대한 방어에 나설 것"이라고 밝혔습니다.

요약문

미국 남부지방법원이 스타벅스 과일 음료에 망고 드래곤푸르트, 파인애플 패션푸르트 등에 실제로 과일이 들어가지 않았다는 이유로 집단소송을 기각했다.

04. 번역_opentest

기사 원문 중 문장

카카오 노동조합 '크루 유니언'은 배임·횡령 혐의로 전 재무그룹장을 경찰에 고발했다고 밝혔습니다.

번역문

Kakao's labor union "Crude Unions" said it filed a complaint with the police for the all-time financial groups on charges of embezzlement and embezzlement.

04. 요약 및 번역_opentest

원문

(로스앤젤레스=연합뉴스) 임미나 특파원 = 세계 최대 엔터테인먼트 기업인 월트디즈니컴퍼니(이하 디즈니)가 향후 10년간 놀이공원(테마파크)과 크루즈 등 사업에 약 80조원을 지출한다는 계획을 발표했다.

디즈니는 19일(현지시간) 미 증권거래위원회(SEC)에 제출한 보고서에서 "디즈니 파크, 체험과 제품(DPEP) 사업 부문에 대한 투자를 확대해 약 10년 동안 해당 부문의 연결 자본 지출을 약 600억달러(약 79조7천400억원)로 늘리겠다"며 "이는 이전 약 10년간의 지출과 비교해 거의 2배 규모"라고 밝혔다.

그러면서 "회사는 신중하고 균형 잡힌 방식으로 자본을 배분한다는 원칙을 견지하면서 강력한 수익을 창출할 것으로 예상되는 프로젝트에 우선순위를 두고 있다"며 "국내외 놀이공원과 크루즈 라인의 수용 능력을 확대하는 데 투자할 것"이라고 설명했다.

이같은 계획은 최근 미디어 환경 변화로 TV·방송 네트워크 사업이 사양 길로 접어든 가운데, 세계적으로 매출이 상승세인 놀이공원·체험형 사업에 집중하려는 전략으로 풀이된다.

지난 3분기 디즈니의 DPEP 사업 부문 매출은 83억달러(약 11조원), 영업이익은 24억달러(약 3조2천억원)로, 작년 동기 대비 각각 13%, 11% 증가했다. 특히 상하이와 홍콩에 있는 디즈니 리조트의 매출 성장세가 두드러졌다.

디즈니는 이날 투자자들을 대상으로 한 행사에서 홍콩, 파리, 도쿄, 상하이 등 미국 외 지역의 테마파크에 애니메이션 '겨울왕국'과 '주토피아'를 주제로 한 놀이기구를 추가하는 방안을 언급했다고 미 경제매체 CNBC는 전했다.

디즈니는 또 디즈니 캐릭터와 마블 슈퍼히어로 등을 활용한 크루즈 사업을 카리브해와 유럽, 호주 등지에서 해왔으며, 지난 4월에는 싱가포르 등 아시아 시장에도 진출한다고 밝힌 바 있다.

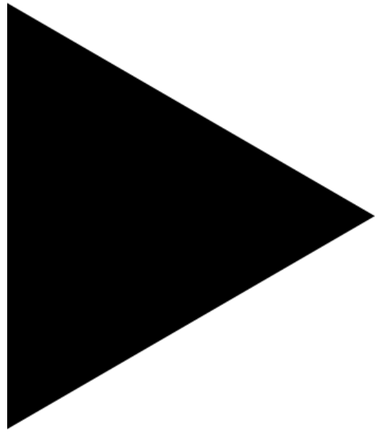
요약문

디즈니는 2020년까지 놀이공원 사업과 크루즈 사업에 약 80조 원을 지출한다는 계획을 발표했고 이를 위해 관련 시설과 장비를 확장한다고 밝혔다

번역문

Disney announced that it will spend about 80 trillion won to play park business and cruise projects by 2020, and it will expand related facilities and equipment to this end.

04. 코드 시연 영상



05. 자체 평가

05. 자체 평가

기대 효과

> 자체적으로 데이터 학습 성공
  > 번역 모델에 다른 언어를 넣어 다문화 사용자를 대상으로 하는 서비스로의 발전 가능성

> 요약과 번역 모두 transformer 기반의 모델이므로 입력 데이터만 변환해주면 다른 도메인으로 모델 구축 가능
- 확장 가능성이 높음

> 향후 모델 경량화 발전 가능 (모델 레이어 축소, Linear 차원 축소 등)

05. 회고

장선

생성 모델 2개, 요약과 기계 번역을 시간과 컴퓨팅 자원이 부족해서 돌릴 때마다 힘들게 학습을 하는 거 같아서
만약에 자원을 생각하지 않고 학습을 할 수 있었다면 더 좋은 퀄리티가 만들어지지 않았을까? 라는 생각이 들었습니다.
프로젝트를 할 때 중요한 건 시간과 개발자의 능력, 활용할 데이터, 환경 및 조건에 맞게 프로젝트를 설정해야한다는 것입니다.
또한 다음으로 나아가는 방향으로는 다른 언어를 지원하게 끔 하고, 모델 성능을 늘리고, 더 많은 모델을 다루면서 더 좋은 모델을 찾아 범용적인 모델을 만들어 개선했으면 좋겠습니다.
생각보다 실수와 업무가 많았고, 수업에 없는 내용들이 실제 개발에서는 많은 기술들이 들어가는 걸 알게 되었습니다. 다들 아는 게 많다고 잘한다고 칭찬해 주셨지만, 제가 오히려 짐 덩이가 된 게 아닌가 싶을 정도로 너무 다들 잘해주셔서 감사하게 생각했고 이번 프로젝트를 디딤돌로 삼아 다음엔 더 잘하도록 노력하겠습니다. 마지막으로 팀원들에게 감사한 말씀 드립니다

석영

먼저 프로젝트에서 모델개발 부분에 집중할 수 있도록 팀원들이 배려해줘서 너무 감사했습니다.
그리고 결과물을 만드는 과정에서 복잡한 모델 구조를 구체적으로 이해하고 설명할 기회가 있어서 행운이었다는 생각이 듭니다.
프로젝트를 끌고 가는 힘과 문제 해결 능력이 월등한 팀원들을 만나 함께 프로젝트를 수행할 수 있어서 좋은 결과물 만들어낼 수 있었고,
최종적으로 성능이 잘 나와서 개인적으로 자신감도 생기는 프로젝트 경험이었습니다.

현호

이번 프로젝트를 통해 문제 해결에 있어 나의 관점으로만 바라보는 것보다 여러 사람의 관점과 함께 입체적으로 바라보는 것이 중요하다고 느꼈습니다.
그리고 빠른 처리 속도와 유의미한 결과를 지향하는 건 중요하지만 관련 개념을 정확히 이해하고 스스로 생각할 수 있는 능력이 더 중요함을 체험할 수 있는 시간이었습니다. 또 앞으로 개발하기 전에, 발생할 수 있는 변수와 여러 관점 및 조언들이 최대한 반영되어 섬세하게 개발을 설계함으로써 리스크를 최소화하여 개발을 진행하는 것이 중요하다는 것을 느꼈습니다. 마지막으로 끝까지 함께 해준 팀원들께 감사의 인사말을 전합니다.

05. 회고

지애

fine-tuning이 가장 아쉬운 부분 중 하나입니다. 레이어를 늘려서 아예 성능을 높여보고도 싶었고, 줄일 때는 어떻게 줄여야 성능이 떨어지지 않을 수 있는지 좀 더 시도해 볼 수 있을 것 같습니다. 우리와 같은 환경에서도 잘 돌아갈만큼 가벼우면서 성능이 좋은 모델들이 많다면 더 많은 사람들이 부담없이 모델을 가공하고 학습시켜 볼 수 있을테니 말입니다.
모델의 구조와, 개념을 이해하려고 공부 했던 시간들이 가장 좋았고, 또 제가 가장 필요한 부분이었던 것 같습니다. 그 과정에서 여러명이 모이니까, 내가 이해하고 있는 부분이 맞는지 확인 받을 수도 있었고, 또 확인해주면서 오히려 이해가 잘 됐던 것 같아서 좋았습니다. 생성모델을 시도해봤다는 그 자체에 의의를 두자고 했지만, 하다보니 성능도, fine-tuning도 욕심이 나서 열심히 했던 과정을 결과적으로 프로젝트에 사용하진 못했지만 모델 뜯어보는 실력이 많이 늘은 것 같아 기분이 좋습니다. 학습할 때, 데이터가 많거나, 배치가 크거나, 에폭을 많이 돌거나 그냥 시도때도 없이 오류나고 끊기는 환경에서도 다들 끝까지 붙잡고 잠도 안자고 돌려본거 정말 고맙습니다. 덕분에 많이 배우는 프로젝트 한 것 같습니다. 다들 수고하셨습니다!

혜인

이번 프로젝트는 이해하지 못했던 딥러닝에 대해 공부할 수 있는 좋은 기회였습니다. 모델에 대한 개념 뿐만 아니라 코드까지 직접 보며 혼자 이해하기 어려웠던 것들도 팀원들에게 배울 수 있어서 좋았습니다. 팀원들에게 많은 도움을 받아 끝까지 할 수 있었습니다.
팀원 모두들 감사합니다:)

은지

인공지능을 공부하는 학생으로서 chat GPT와 같은 생성모델을 직접 학습시키고 만들어 보면서 많은 점을 배울 수 있었습니다.
트랜스포머와 같이 시퀀스 데이터 처리에 핵심적인 모델의 구조와 실제 구현 방식을 실습해보고 BART와 같은 최신 모델을 우리 데이터에 맞게 학습시키고 모델 레이어를 다양하게 조작해보며 모델의 이해도를 높일 수 있었습니다. 다만, 많은 양의 학습데이터 구축과 GPU 파워가 핵심인 생성모델을 구현하는데 컴퓨팅 파워가 아쉬워 많은 컴퓨팅을 활용했다면 더 나은 결과를 도출해낼 수 있지 않았을까 아쉬움도 듭니다.
마지막으로 모델 개발에 전념할 수 있도록 프론트와 백을 담당해주신 팀원 분께 감사드립니다.

06. Q & A

감사합니다!